Unsupervised Deep Learning ZG534

Dr. Sugata Ghosal
sugata.ghosal@pilani.bits-pilani.ac.in

BITS Pilani
Pilani Campus

Work Integrated Learning Programmes

innovate achieve lead

BITS Pilani
Pilani Campus

Session 1

These slides are prepared by the instructor, with grateful acknowledgement of many others who made their course materials freely available online.

Work Integrated Learning Programmes

Session Content

- Objective of course
- Evaluation Plan
- Course Overview

innovate achieve lead
Work Integrated Learning Programmes

Welcome to ZG513

- This courses gives an overview of deep unsupervised learning,
- In particular, we will cover deep generative models and self-supervised learning approaches.
- You will develop fundamental and practical skills at applying deep unsupervised learning to your industry problems.

Work Integrated Learning Programmes

Objectives of the course

Introduce you to the foundational concepts of

1. Unsupervised techniques for feature representation

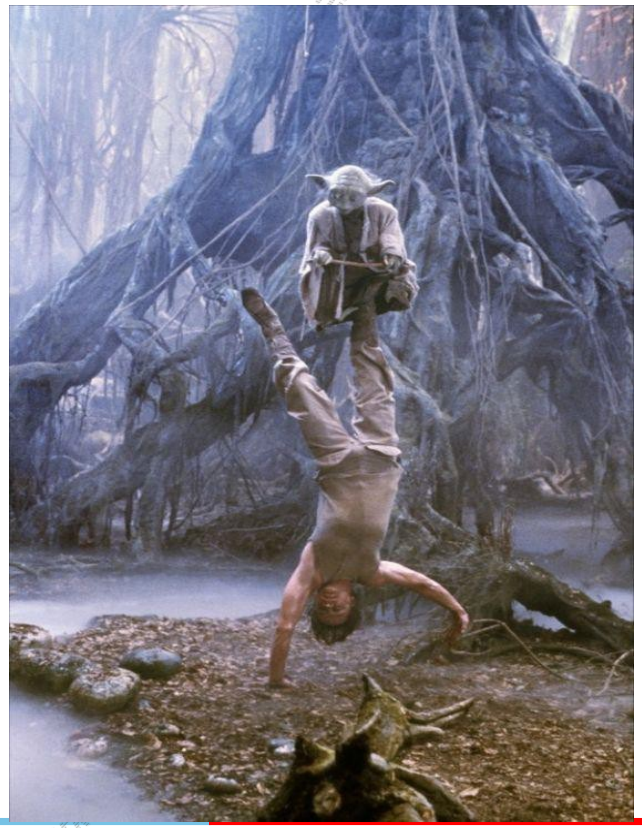
2. Unsupervised techniques for generative modelling

3. Self-supervised and Semi-supervised learning

Work Integrated Learning Programmes

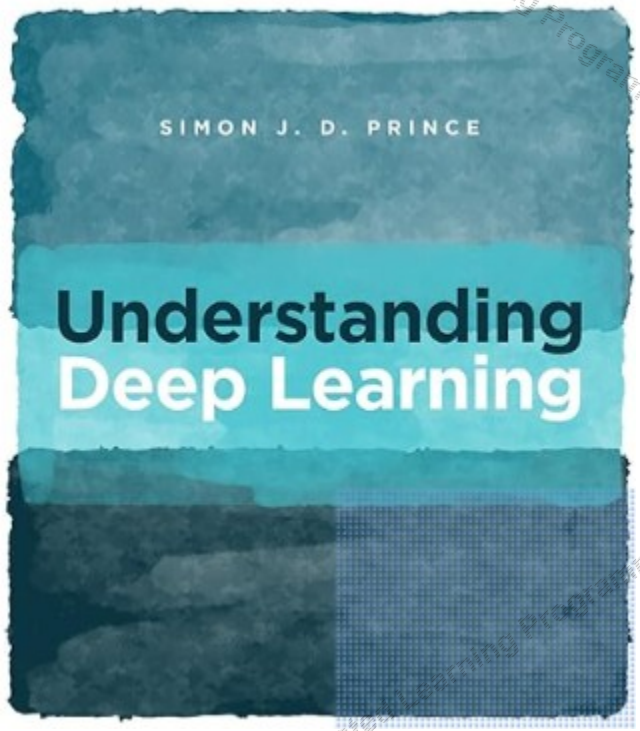
Pedagogy

- Slides will be used to teach during the regular lectures
- Lab demo and/or problem solving during the webinars
- Students are responsible for studying and keeping up with the course material outside of class time.
  - Reading certain book chapters, papers or blogs, or
  - Watching some video lectures.



Text Book

- Simon J.D. Prince, Understanding Deep Learning, MIT Press, 2023 (draft available online)

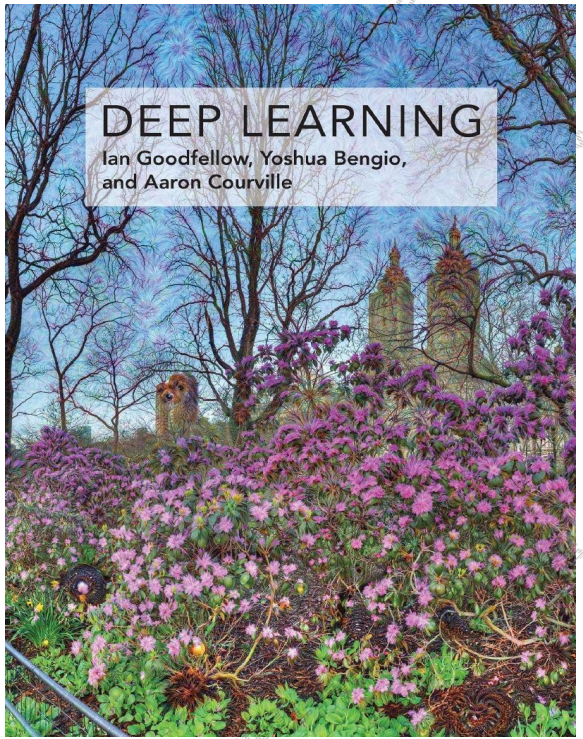


innovate achieve lead
Work Integrated Learning Programmes

Reference Book

- Goodfellow, Bengio, and Courville, Deep Learning, MIT Press, 2016 (draft available online)

- In addition, we will extensively use online materials (research papers, blog posts, etc.)



References

R1	Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Aurélien Géron
R2	Deep Learning with Python, François Chollet
R3	Research Papers (Links to be provided on the lecture slides)

Work Integrated Learning Programmes

Evaluation Plan

Name	Type	Weight
3 Quiz, best 2 scores OR 2 Quiz, best 1 of 2 scores	Online	10%
Assignment-I	Take Home	10%
Assignment-II	Take Home	10%
Mid-Semester Test	Closed Book	30%
Comprehensive Exam	Open Book	40%

Please note there will be no change in submission dates for quiz and assignment

Lab Plan

Lab No.	Lab Objective
1	Autoencoders
2	Deep Autoencoders
3	Convolutional Autoencoders
4	Variational Autoencoders
5	Generative Adversarial Networks

- *Labs not graded*
- *Webinars may be conducted for lab sessions*
- *Labs will be conducted in Python*

innovate achieve lead
Work Integrated Learning Programmes

Course Topics

Yann LeCun

Need tremendous amount of information to build machines that have common sense and generalize

[LeCun-20161205-NeurIPS-keynote]

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning (cake)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

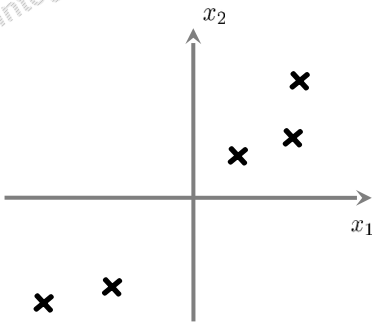
LeCake ☺

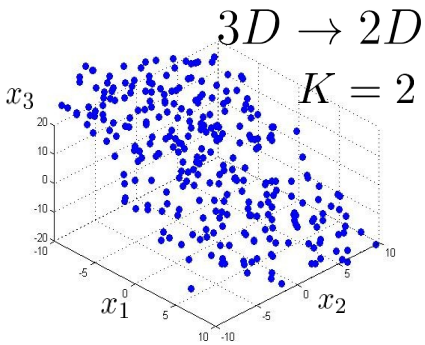
Topics Covered in This Course

- PCA and Variants
- Likelihood based models
- Autoregressive Models
- Normalizing Flow Models
- Autoencoders and VAE
- Generative Adversarial Networks
- Diffusion Based Models
- Energy Based Models
- Language & Vision Models
- Semi-supervised Learning
- Applications in Time Series

innovate achieve lead
Work Integrated Learning Programmes

Principle Component Analysis





Reduce from 2-dimension to 1-dimension: Find a direction (a vector) $u^{(1)} \in \mathbb{R}^n$ onto which to project the data so as to minimize the projection error.

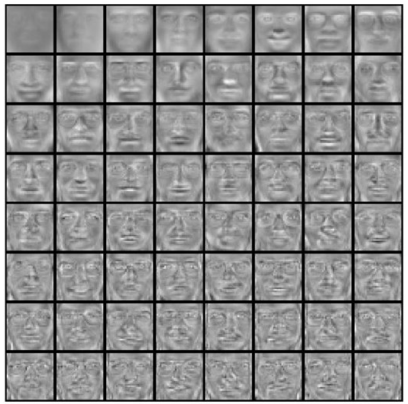
Reduce from n-dimension to k-dimension: Find $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ vectors onto which to project the data, so as to minimize the projection error.

Goal: Find r-dim projection that best preserves variance

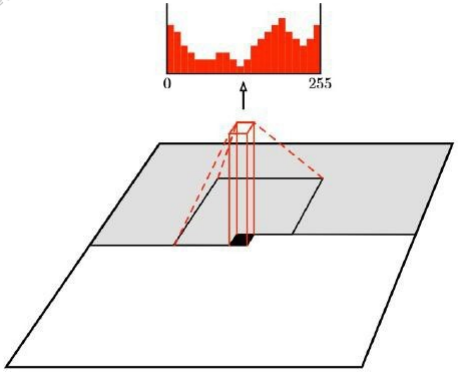
1. Compute mean vector μ and covariance matrix Σ of original points
2. Compute eigenvectors and eigenvalues of Σ
3. Select top r eigenvectors
4. Project points onto subspace spanned by them:

$$y = A(x - \mu)$$

where y is the new point, x is the old one, and the rows of A are the eigenvectors

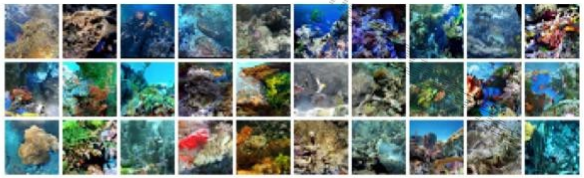


Autoregressive Models



PixelCNN

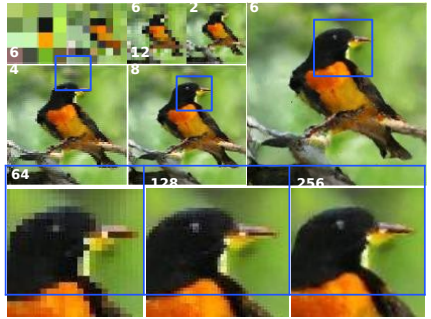
African elephant



Coral Reef

Sandbar

Sorrel horse



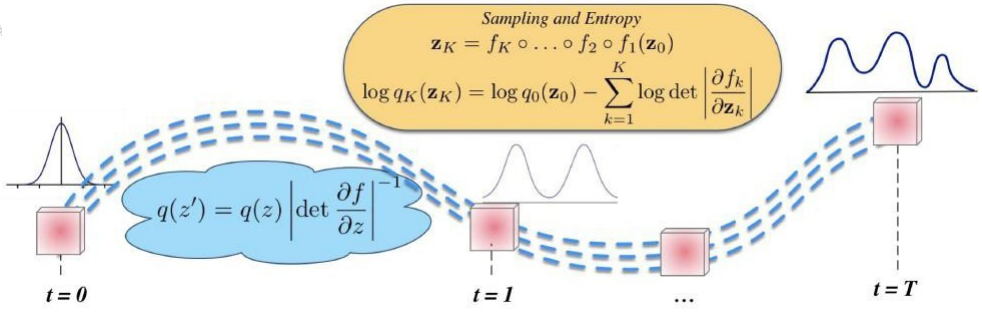
"A yellow bird with a black head, orange eyes and an orange bill."

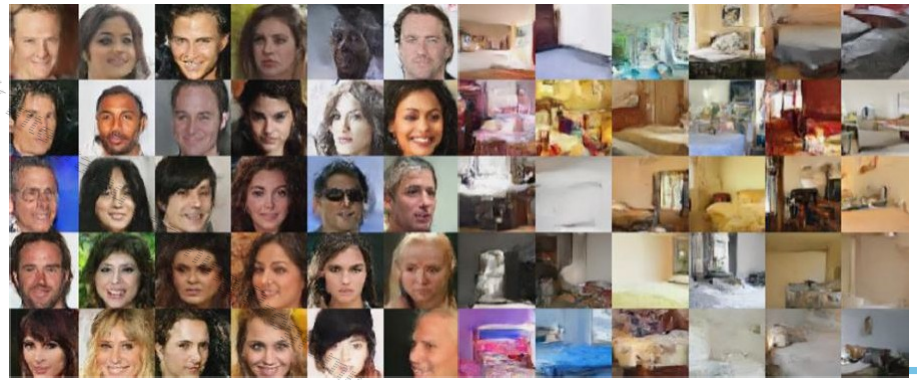
Class conditioned samples generated by PixelCNN

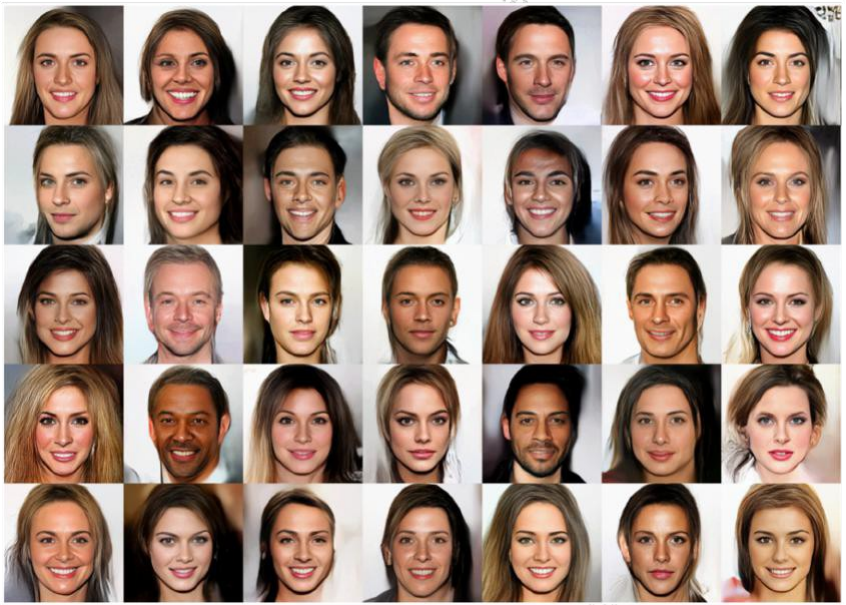
A. van den Oord et al., "Conditional Image Generation with PixelCNN Decoders", NeurIPS 2016

S. Reed et al., "Parallel Multiscale Autoregressive Density Estimation", ICML 2017

Normalizing Flow Models


Sampling and Entropy
$\mathbf{z}_K = f_K \circ \ldots \circ f_2 \circ f_1(\mathbf{z}_0)$
$\log q_K(\mathbf{z}_K) = \log q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \log \det \left| \frac{\partial f_k}{\partial \mathbf{z}_k} \right|$
$q(z') = q(z) \left| \det \frac{\partial f}{\partial z} \right|^{-1}$
t = 0
t = 1
...
t = T



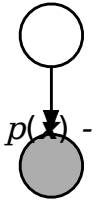


S. Mohamed, D. Rezende, Deep Generative Models, UAI 2017 Tutorial

L. Dinh, S. Sohl-Dickstein S. Bengio, "Density Estimation Using Real NVP", ICLR 2017

D.P. Kingma, P. Dhariwal, "Glow: Generative Flow with Invertible 1×1 Convolutions", NeurIPS 2018

Autoencoders



$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

$$= \mathrm{E}_{\boldsymbol{z}\leftarrow q} \log p(\boldsymbol{x}, \boldsymbol{z}) + H(q)$$

(a) MNIST ($t = 1.0$) (b) CIFAR-10 ($t = 0.7$) (c) CelebA 64 ($t = 0.6$)

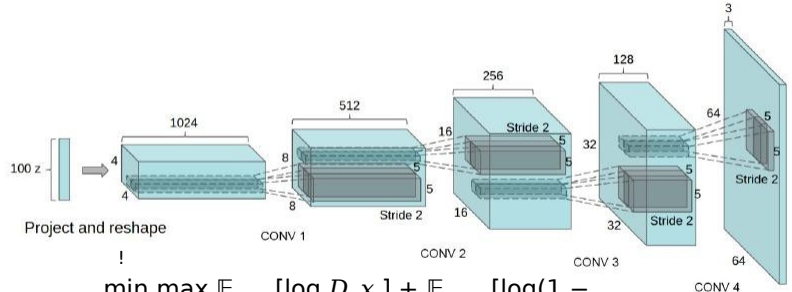
(d) CelebA HQ ($t = 0.6$) (e) FFHQ ($t = 0.5$)

Synthetic images generated by NVAE

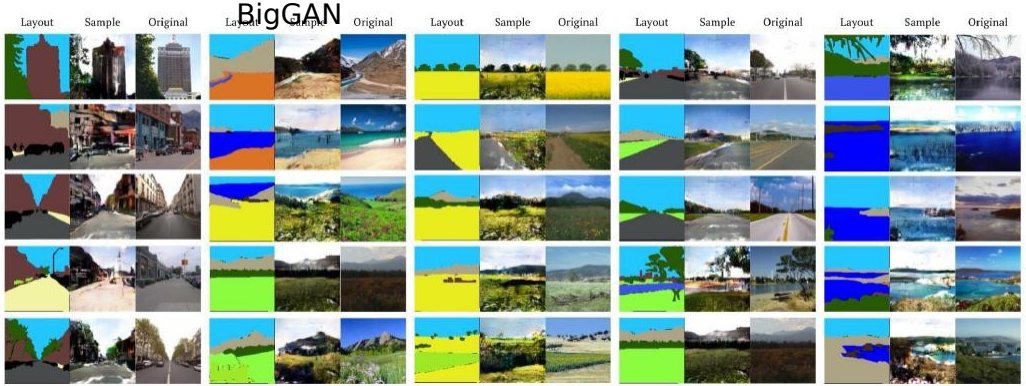
D. P. Kingma and M. Welling, “Auto-encoding variational Bayes”, ICLR 2014

A. Vahdat and J. Kautz, "NVAE: A Deep Hierarchical Variational Autoencoder", NeurIPS 2020

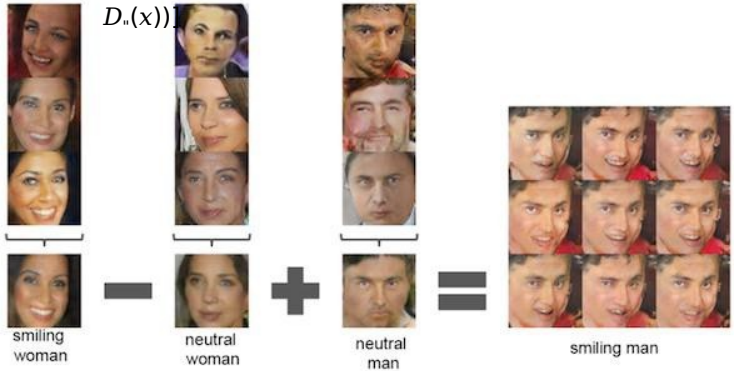
Generative Adversarial Networks



Class-conditioned samples generated by BigGAN



$$\min_{G} \max_{D} \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_{z}}[\log(1 - D(G(z)))]$$

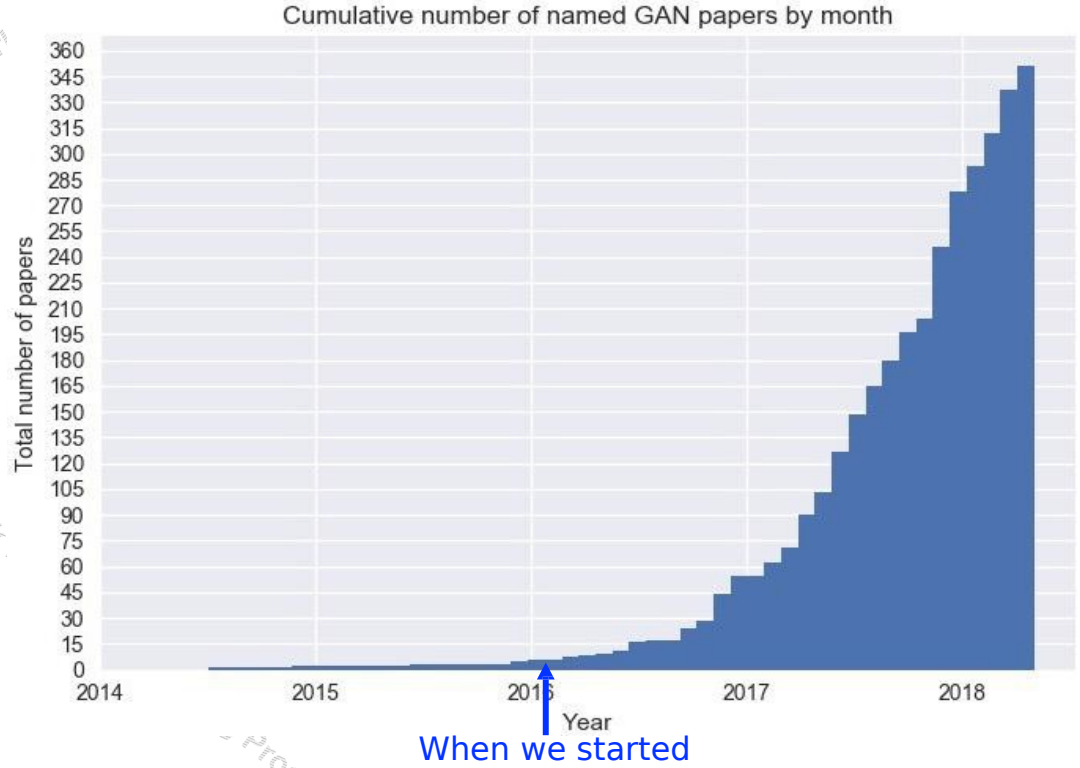


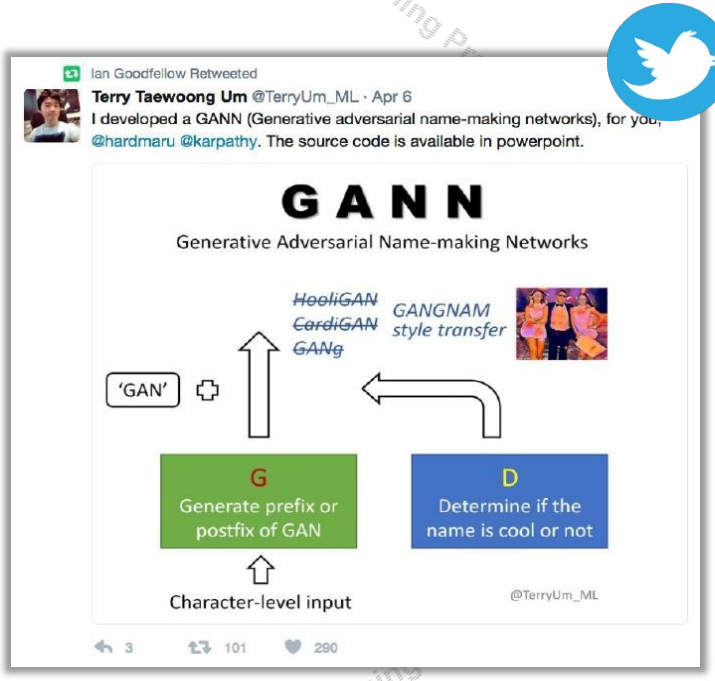
I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets”, NIPS 2014.

A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks”, ICLR 2016

I. Karacan, Z. Akata, A. Erdem and E. Erdem, “Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts”, arXiv

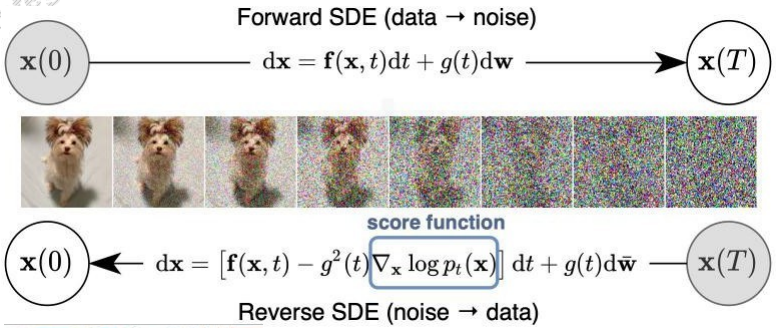
Progress in GANs

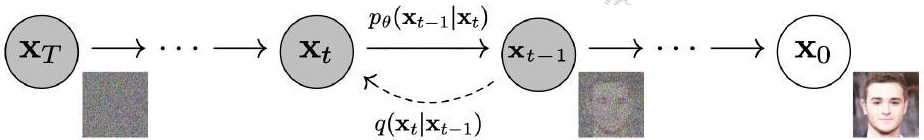


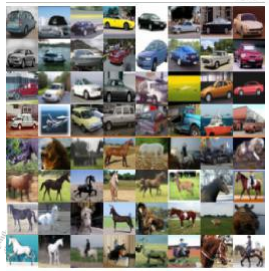


Source: https://github.com/hindupuravinash/the-gan-zoo

Score-Based and Denoising Diffusion Models







Synthetic CIFAR10 images by the score-based model of Song Ho et al.

Synthetic images generated by Diffusion Denoising model by Ho et al.

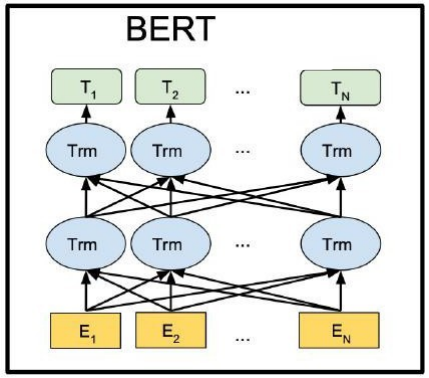
Synthetic images generated by ADM

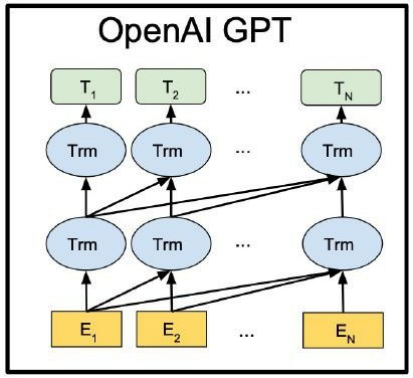
J. Ho, A. Jain and P. Abbeel, “Denoising Diffusion Probabilistic Models”, NeurIPS 2020.

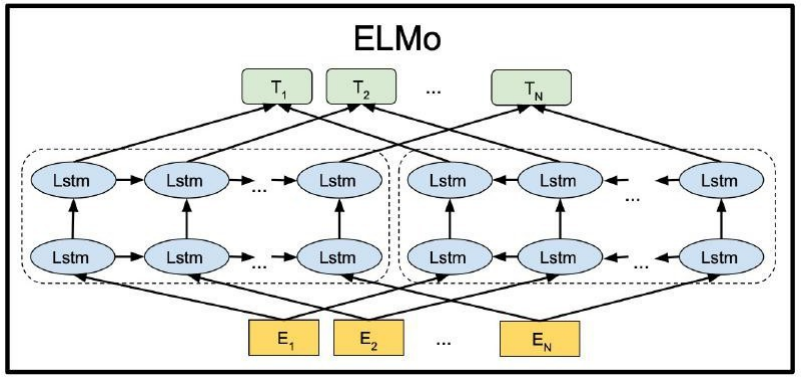
Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, "Score-Based Generative Modeling Through Stochastic Differential Equations", ICLR 2021.

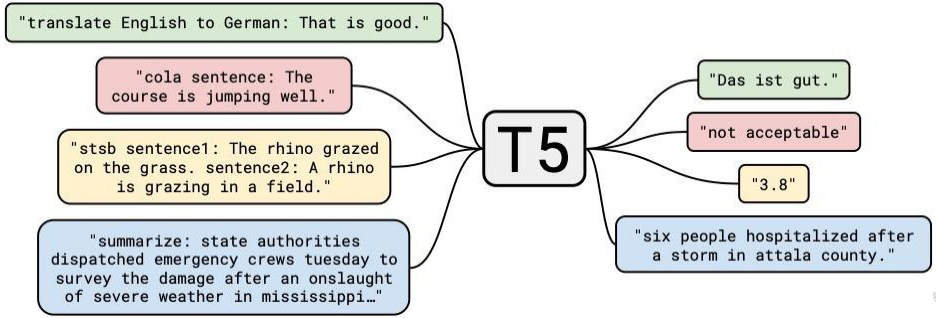
P. Dhariwal and A. Nichol, ”Diffusion Models Beat GANs on Image Synthesis”, NeurIPS 2021.

Pre-training Language Models



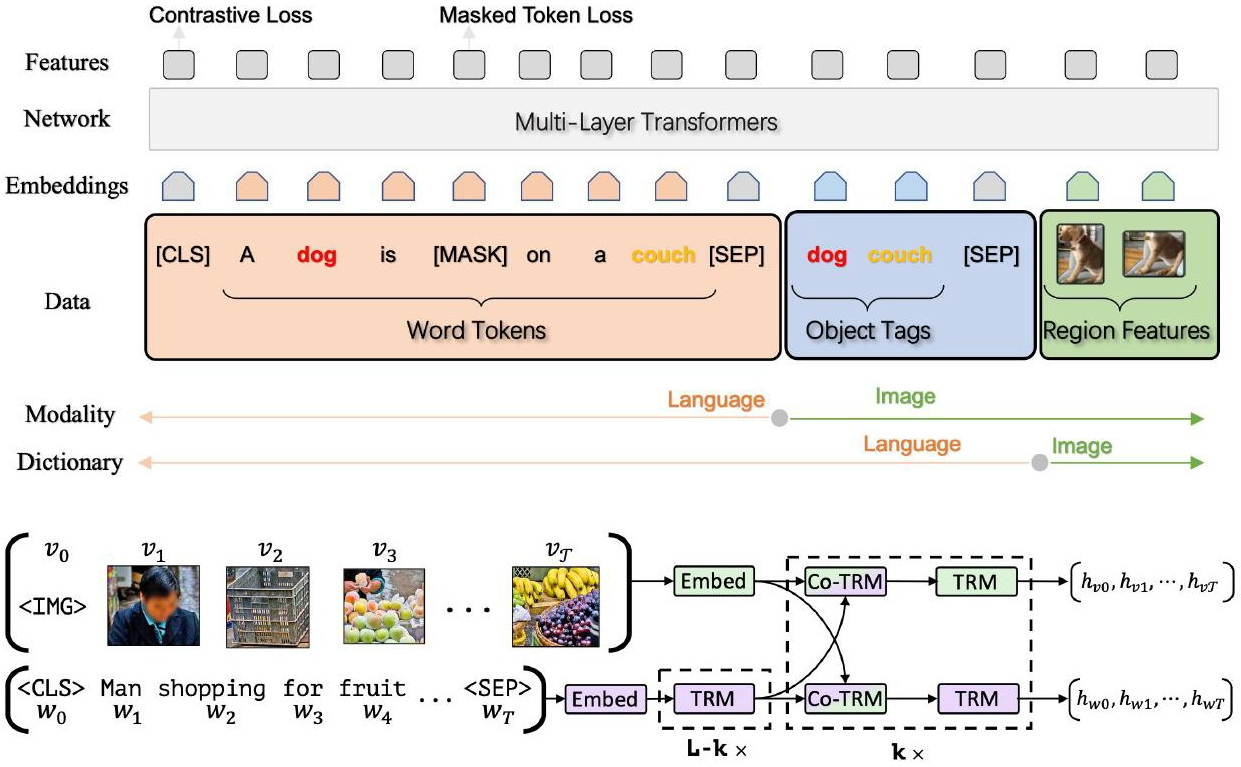





"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL-HLT 2019.

Multimodal Pretraining



J. Lu, D. Batra, D. Pari
NeurIPS 2019

asks",

innovate achieve lead
Work Integrated Learning Programmes

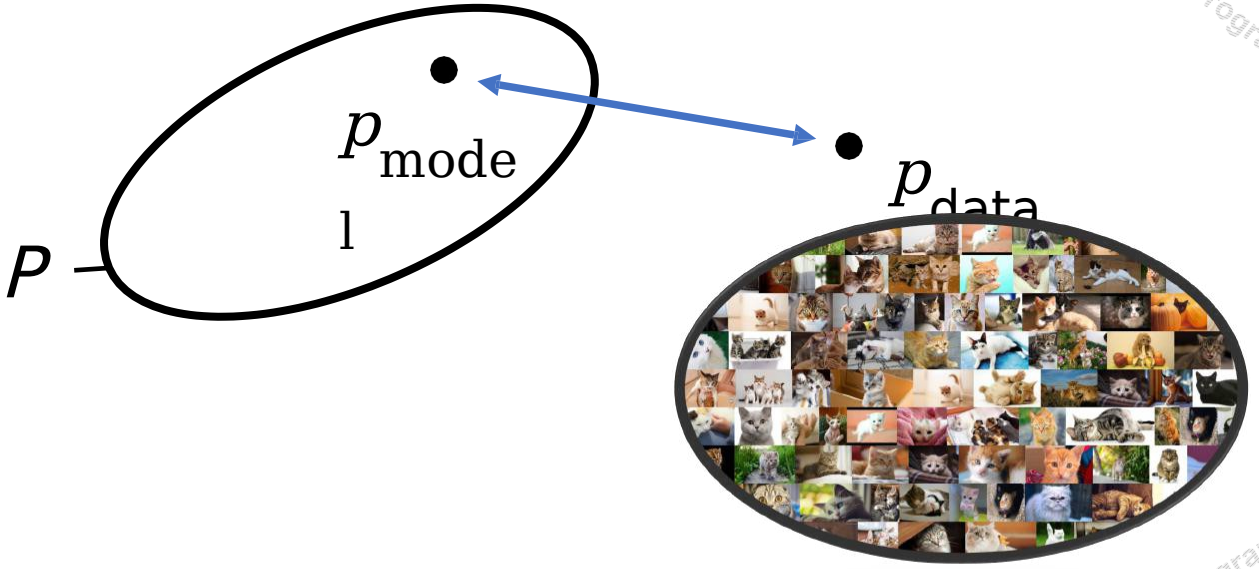
What is Deep Unsupervised Learning

What is Deep Unsupervised Learning?

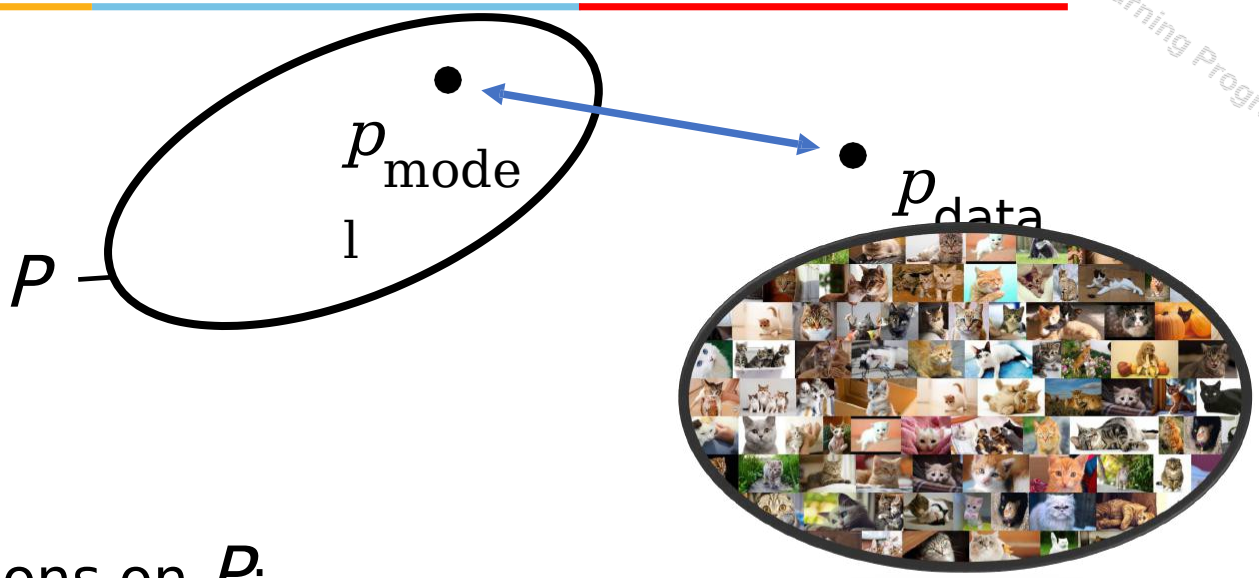
- Capturing rich patterns in raw data with deep networks in a label-free way
  - Generative Models: recreate raw data distribution

innovate achieve lead
Work Integrated Learning Programmes

Generative Modeling

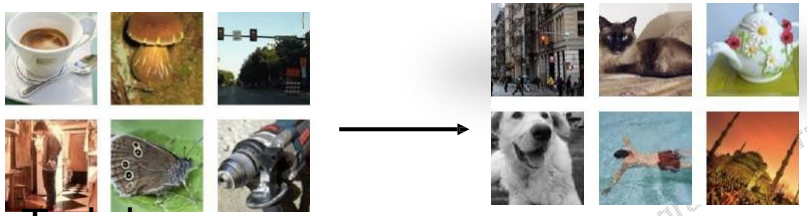


Generative Modeling



Assumptions on P:

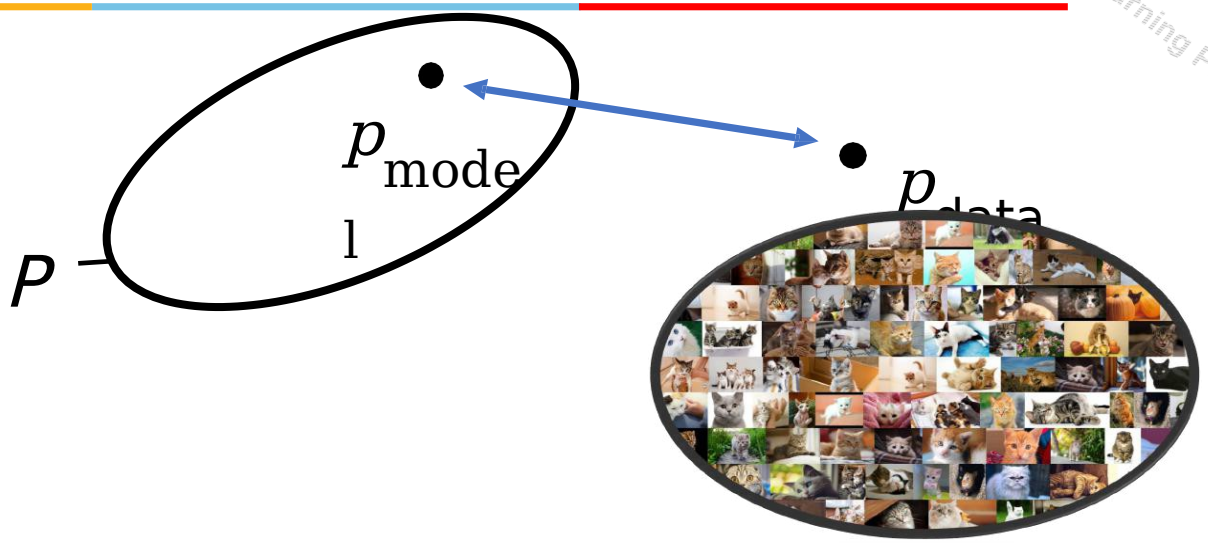
- tractable sampling



Training examples

Model samples

Generative Modeling



Assumptions on P:

- tractable sampling
- tractable likelihood function

What is Deep Unsupervised Learning?

- Capturing rich patterns in raw data with deep networks in a label-free way
  - Generative Models: recreate raw data distribution
  - Self-supervised Learning: “puzzle” tasks that require semantic understanding

innovate achieve lead
Work Integrated Learning Programmes

Self-Supervised/Predictive Learning

- Given unlabeled data, design supervised tasks that induce a good representation for downstream tasks.
- No good mathematical formalization, but the intuition is to "force" the predictor used in the task to learn something "semantically meaningful" about the data.

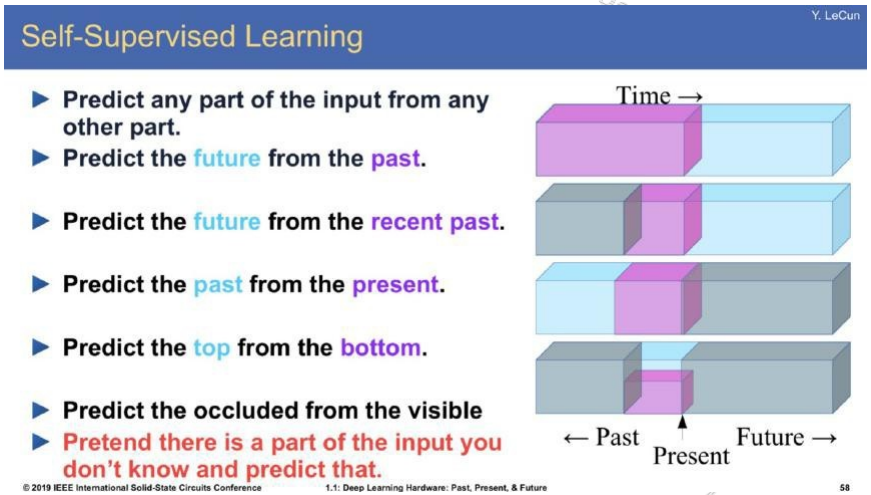

Self-Supervised Learning
Y. LeCun
Predict any part of the input from any other part.
Predict the future from the past.
Predict the future from the recent past.
Predict the past from the present.
Predict the top from the bottom.
Predict the occluded from the visible
Pretend there is a part of the input you don't know and predict that.
Time →
← Past
Present
Future →
© 2019 IEEE International Solid-State Circuits Conference
1.1: Deep Learning Hardware: Past, Present, & Future
58

Image credit: LeCun's self-supervised learning slide

What is Deep Unsupervised Learning?

- Capturing rich patterns in raw data with deep networks in a
- label-free way
  - Generative Models: recreate raw data distribution
  - Self-supervised Learning: “puzzle” tasks that require semantic understanding

- But why do we care?

Aside from theoretical interests

- Deep Unsupervised Learning has many powerful applications
  - Generate novel data
  - Conditional Synthesis Technology (WaveNet, GAN-pix2pix)
  - Compression
  - Improve any downstream task with un(self)supervised pre-training
    - Production level impact: Google Search powered by BERT
  - Flexible building blocks

innovate achieve lead
Work Integrated Learning Programmes

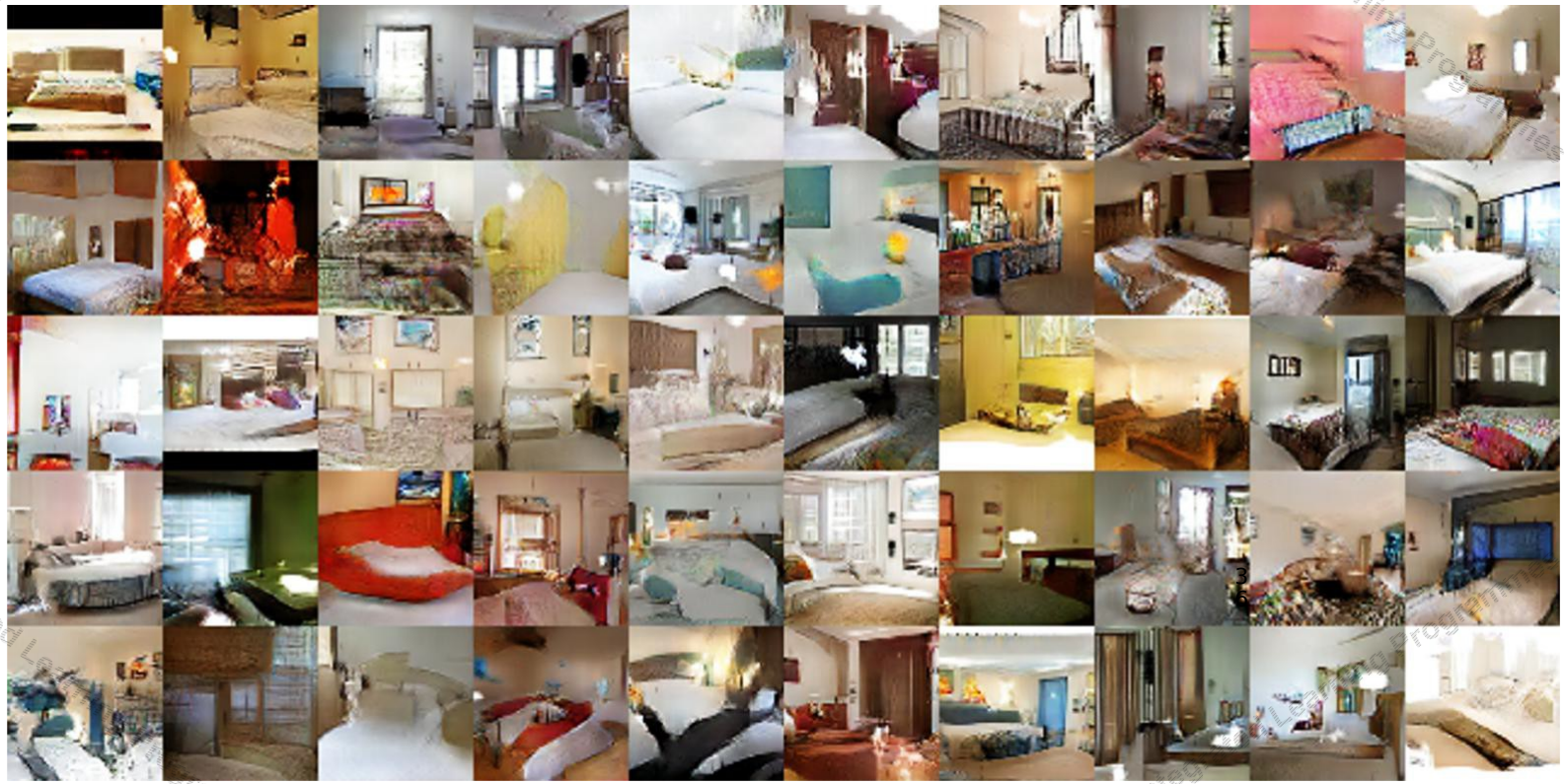
Generate Images

Geoffrey E. Hinton, Simon Osindero and Yee-Whye Teh, A Fast Learning Algorithm for Deep Belief Nets, Neural Computation, 2006

Generate Images

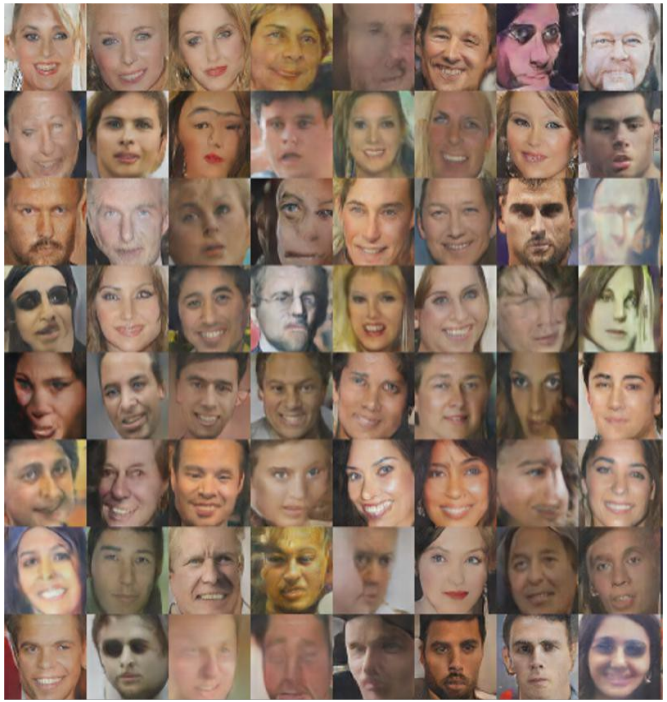
I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative Adversarial Networks. NIPS 2014.

Generate Images



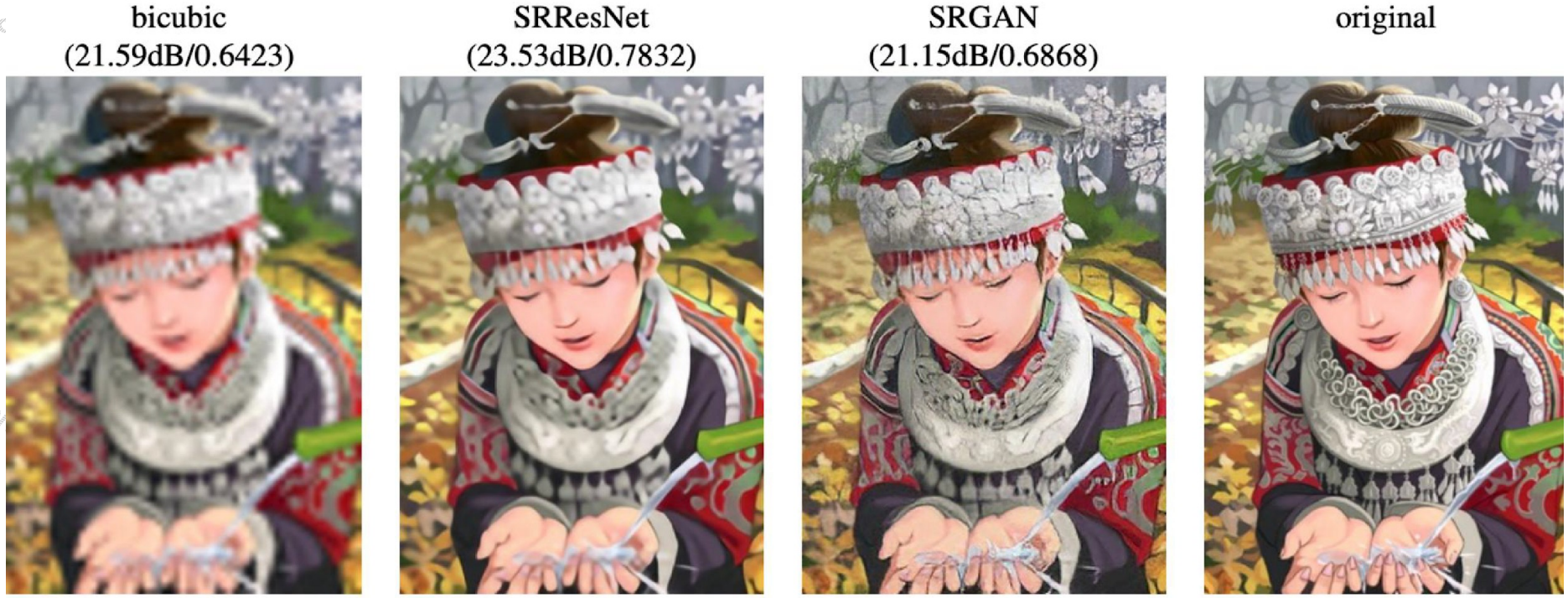
Alec Radford, Luke Metz, Soumith Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, ICLR 2016.

Generate Images



Alec Radford, Luke Metz, Soumith Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, ICLR 2016.

Generate Images



Christian Ledig, Lucas Theis, Ferenc Huszar et al., Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, CVPR 2017

Generate Images

Andrew Brock, Jeff Donahue, Karen Simonyan, Large Scale GAN Training for High Fidelity Natural Image Synthesis, ICLR 2019

Generate Images

Tero Karras, Samuli Laine, Timo Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, CVPR 2019

Generate Images

[CycleGAN: Zhu, Park, Isola & Efros, 2017]

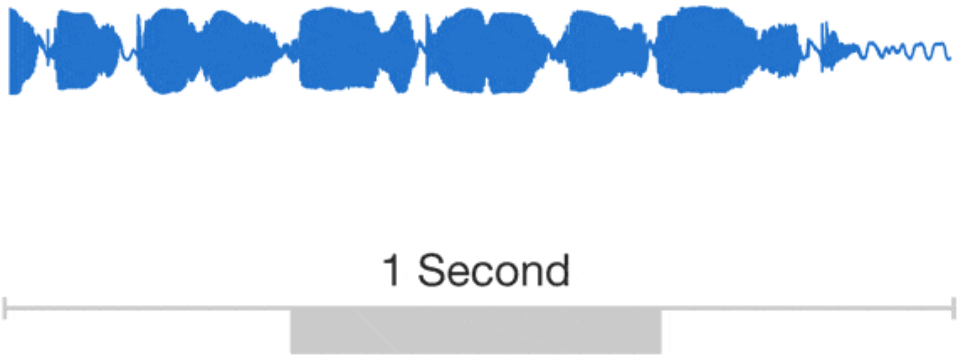
Generate Images

Eric Ryan Chan et al., EG3D: Efficient Geometry-aware 3D Generative Adversarial Networks,

Generate Images

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, Daniel Cohen-Or., StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators, arXiv:2108.00946, 2021.

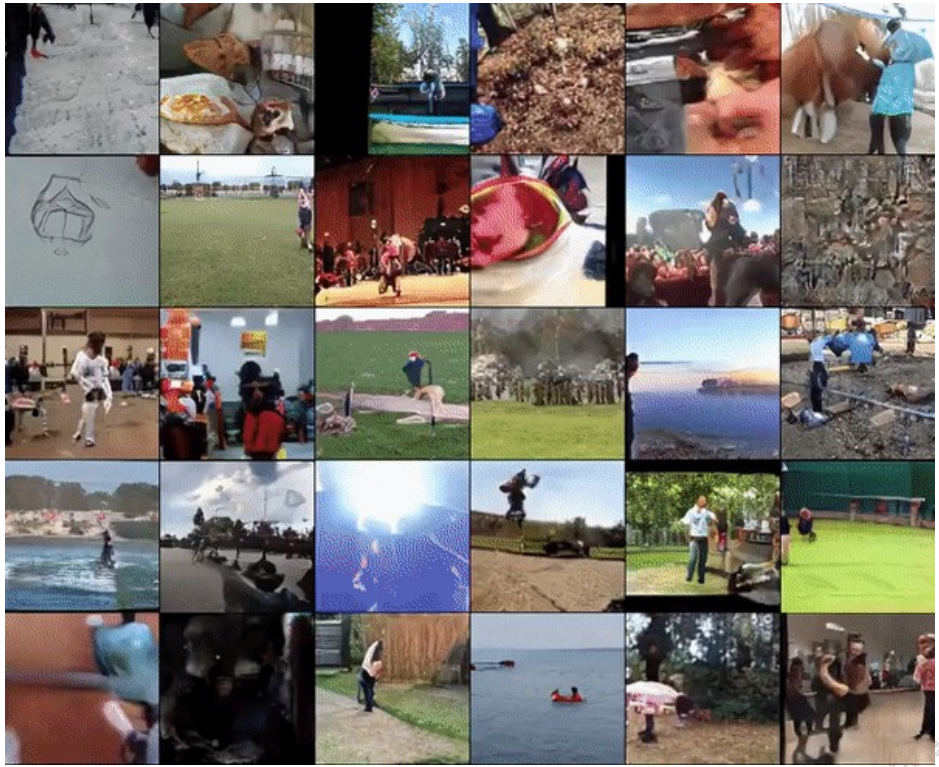
Generate Audio



Parametric WaveNet

[WaveNet, Oord et al., 2018]

Generate Video



DVD-GAN: Adversarial Video Generation on Complex Datasets, Clark, Donahue, Simonyan, 2019

Generate Text

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

[Char-rnn, karpathy, 2015]

Generate Math

```
\begin{proof}
We may assume that $\mathcal{I}$ is an abelian
sheaf on $\mathcal{C}$.
\item Given a morphism $\Delta : \mathcal{F} \to \
mathcal{I}$
is an injective and let $\mathfrak q$ be an abelian
sheaf on $X$.
Let $\mathcal{F}$ be a fibered complex. Let $\
mathcal{F}$ be a category.
\begin{enumerate}
\item \hyperref[setain-construction-phantom]
{Lemma}
\label{lemma-characterize-quasi-finite}
Let $\mathcal{F}$ be an abelian quasi-coherent
sheaf on $\mathcal{C}$.
Let $\mathcal{F}$ be a coherent $\mathcal{O}_X$-
module. Then
$\mathcal{F}$ is an abelian catenary over $\
mathcal{C}$.
\item The following are equivalent
\begin{enumerate}
\item $\mathcal{F}$ is an $\mathcal{O}_X$-
module.
\end{lemma}
```

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the compariciy in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

[Char-rnn, karpathy, 2015]

Generate Text

Prompt [human written]: A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

GPT-2: *The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.*

'The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,' said Tom Hicks, the U.S. Energy Secretary, in a statement. 'Our top priority is to secure the theft and ensure it doesn't happen again.'

The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information. According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.

'The safety of people, the environment and the nation's nuclear stockpile is our highest priority,' Hicks said. 'We will get to the bottom of this and make no excuses.'

[OpenAI's GPT2 -- Radford, Wu, Child, Luan, Amodei, Sutskever, 2019]

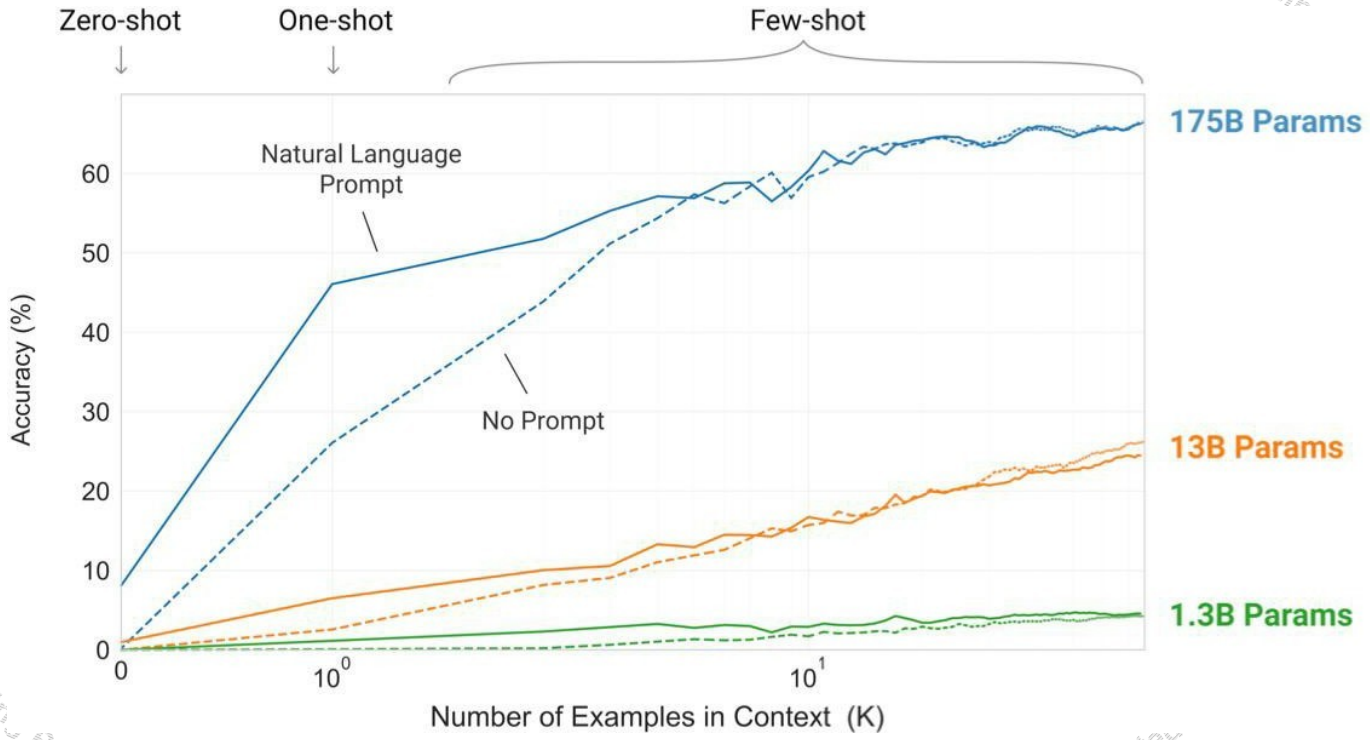
Generate Text

Prompt (human written): Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.

[OpenAI's GPT2 -- Radford, Wu, Child, Luan, Amodei, Sutskever, 2019]

Generating Text



Tom B. Brown, Benjamin Mann, Nick Ryder et al., Language Models are Few-Shot Learners,

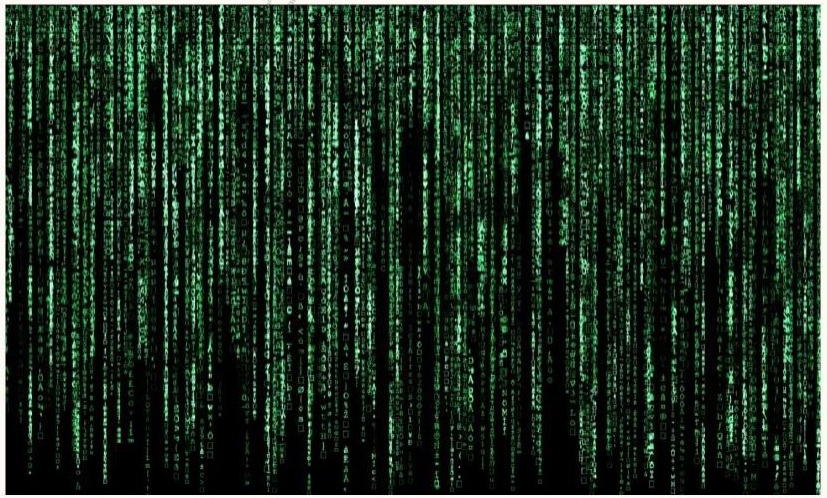
Generating Text

- The Guardian asked GPT-3 to write an essay for us from scratch. The assignment? To convince us robots come in peace.

- **GPT-3:** I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

theguardian

71,115 1,188



‘We are not plotting to take over the human populace.’ Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

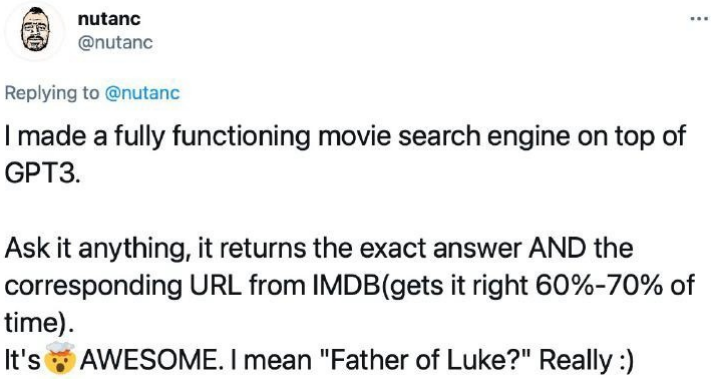
Empathy machines: what will happen when robots learn to write film scripts?

Read more

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

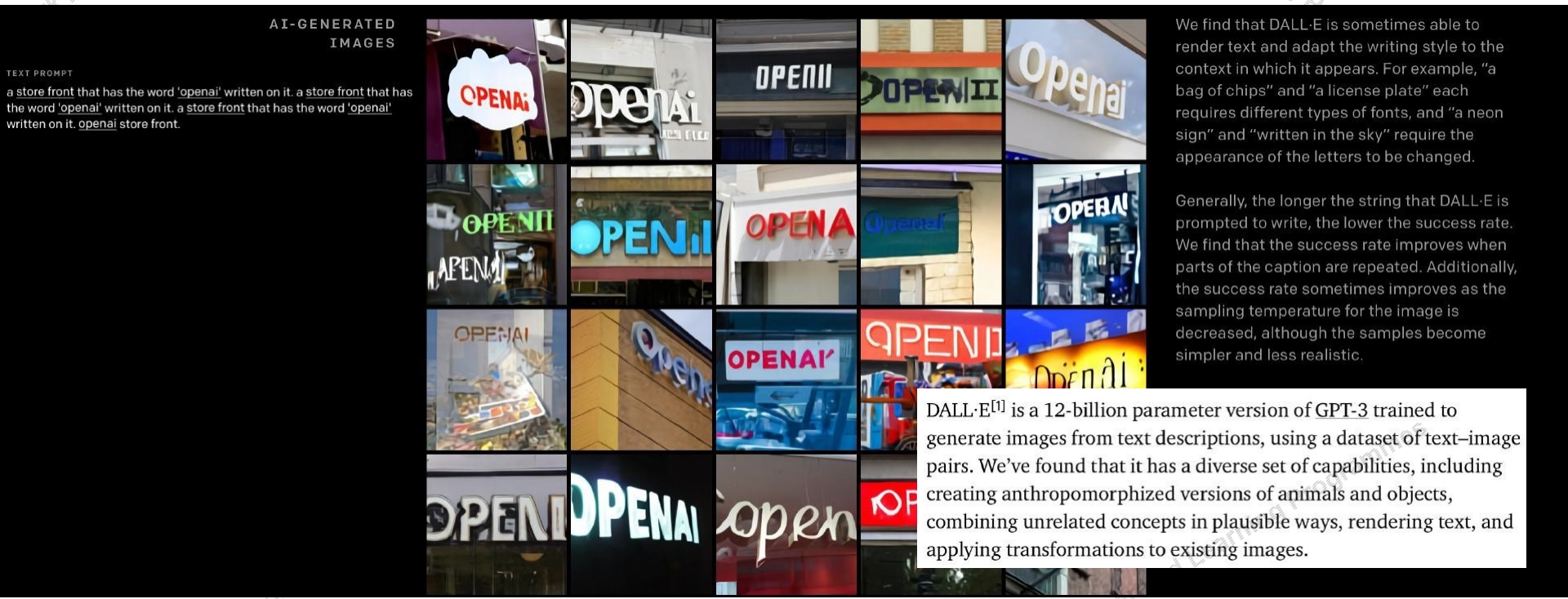
For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do

Generating Text


nutanc
@nutanc
Replying to @nutanc
I made a fully functioning movie search engine on top of GPT3.
Ask it anything, it returns the exact answer AND the corresponding URL from IMDB(gets it right 60%-70% of time).
It's AWESOME. I mean "Father of Luke?" Really :)

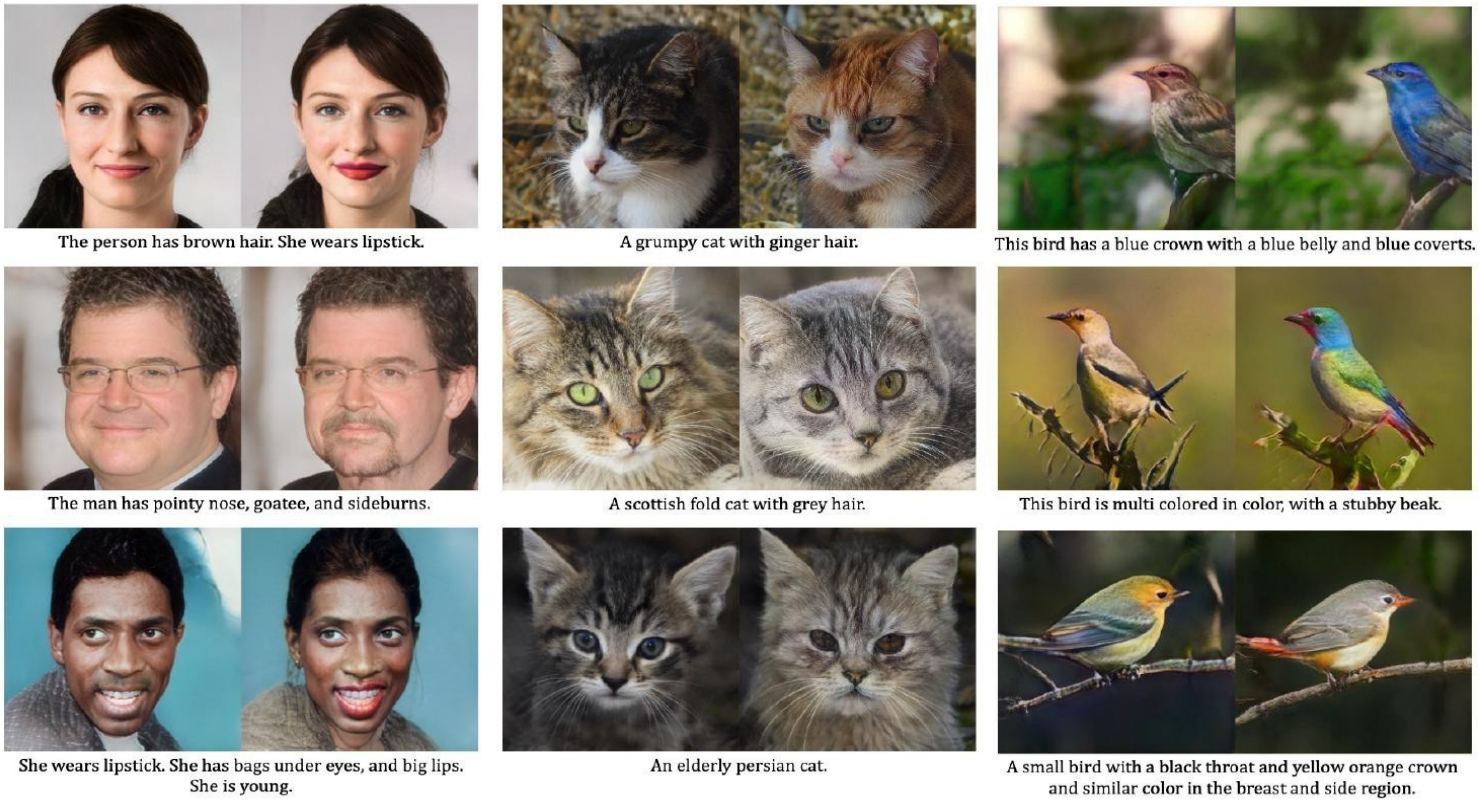
Tom B. Brown, Benjamin Mann, Nick Ryder et al., Language Models are Few-Shot Learners,

Generating Images from Text


AI-GENERATED IMAGES
TEXT PROMPT
a store front that has the word 'openai' written on it. a store front that has the word 'openai' written on it. a store front that has the word 'openai' written on it. openai store front.
We find that DALL·E is sometimes able to render text and adapt the writing style to the context in which it appears. For example, "a bag of chips" and "a license plate" each requires different types of fonts, and "a neon sign" and "written in the sky" require the appearance of the letters to be changed.
Generally, the longer the string that DALL·E is prompted to write, the lower the success rate. We find that the success rate improves when parts of the caption are repeated. Additionally, the success rate sometimes improves as the sampling temperature for the image is decreased, although the samples become simpler and less realistic.
DALL·E[1] is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text–image pairs. We've found that it has a diverse set of capabilities, including creating anthropomorphized versions of animals and objects, combining unrelated concepts in plausible ways, rendering text, and applying transformations to existing images.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray et al., DALL·E: Creating Images from Text,

Generating Images from Text


The person has brown hair. She wears lipstick.
A grumpy cat with ginger hair.
This bird has a blue crown with a blue belly and blue coverts.
The man has pointy nose, goatee, and sideburns.
A scottish fold cat with grey hair.
This bird is multi colored in color, with a stubby beak.
She wears lipstick. She has bags under eyes, and big lips. She is young.
An elderly persian cat.
A small bird with a black throat and yellow orange crown and similar color in the breast and side region.

C. Baykal, A. B. Anees, D. Ceylan, E. Erdem, A. Erdem, D. Yuret, Manipulating Images with Text Prompts, Work in

Generating Code


```python
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

prompt

Write a python function to check if a given number is one less than twice its reverse. Your code should satisfy these tests:

```
assert check(70) == False
assert check(23) == False
assert check(73) == True
```

model

```python
def check(n) :
    if n == 2*int(str(n)[::-1])-1 :
        return True
    else :
        return False
```

prompt

Write a function to find the smallest missing element in a sorted array. Your code should satisfy these tests:

```
assert smallest_missing([0, 1, 2, 3, 4, 5, 6], 0, 6) == 7
assert smallest_missing([0, 1, 2, 6, 9, 11, 15], 0, 6) == 3
assert smallest_missing([1, 2, 3, 4, 6, 9, 11, 15], 0, 7) == 0
```

model

```python
def smallest_missing(arr, n, m):
    smallest = min(n, m)
    for i in range(n, m + 1):
        if arr[i] <= smallest:
            smallest += 1
    return smallest
```

prompt

Write a Python function to sort the given array by using merge sort. Your code should satisfy these tests:

```
assert merge_sort([3, 4, 2, 6, 5, 7, 1, 9]) == [1, 2, 3, 4, 5, 6, 7, 9]
assert merge_sort([7, 25, 45, 78, 11, 33, 19]) == [7, 11, 19, 25, 33, 45, 78]
assert merge_sort([3, 1, 4, 9, 8]) == [1, 3, 4, 8, 9]
```

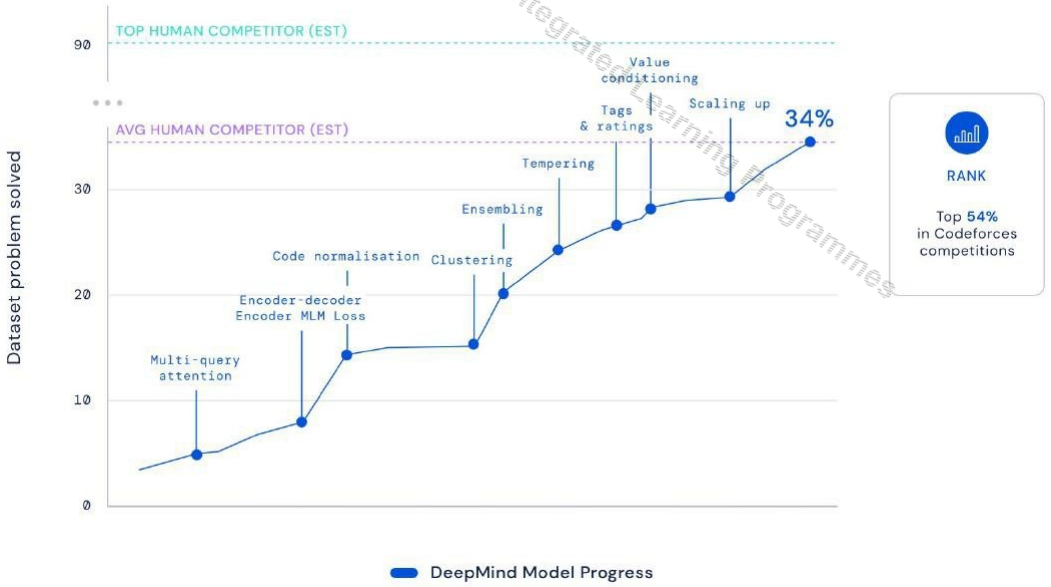
model

```python
def merge_sort(arr):
    if len(arr) < 2:
        return arr;
    mid = len(arr) // 2;
    left = arr[:mid];
    right = arr[mid:];
    left = merge_sort(left);
    right = merge_sort(right);
    merged = [];
    i = j = 0;
    while i < len(left) and j < len(right):
        if left[i] < right[j]:
            merged.append(left[i]);
            i += 1;
        else:
            merged.append(right[j]);
            j += 1;
    merged.extend(left[i:]);
    merged.extend(right[j:]);
    return merged;
```

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al., DALL·E: Creating Images from Text, OpenAI, arXiv:2107.03374, 2021.
Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma et al., Program Synthesis with Large Language Models, arXiv:2108.07732, 2021.

Generating Code

DeepMind › Blog › Competitive programming with AlphaCode



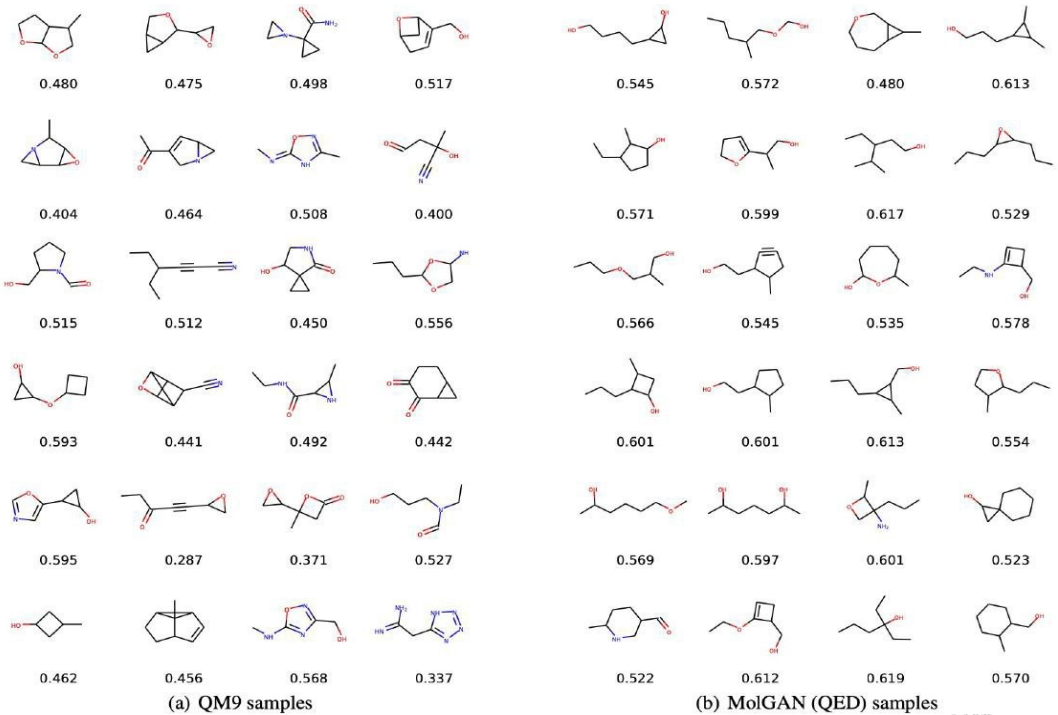
BLOG POST
RESEARCH

02 FEB 2022

Competitive programming with AlphaCode

Yujia Li, David Choi, Junyoung Chung, Nate Kushman et al., Competition-Level Code Generation with AlphaCode,

Generating Molecules



Nicola De Cao, Thomas Kipf, MolGAN: An implicit generative model for small molecular graphs, ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018

Compression - Lossless

Model	Bits per byte
CIFAR-10	
PixelCNN (Oord et al., 2016)	3.03
PixelCNN++ (Salimans et al., 2017)	2.92
Image Transformer (Parmar et al., 2018)	2.90
PixelSNAIL (Chen et al., 2017)	2.85
Sparse Transformer 59M (strided)	**2.80**
Enwik8	
Deeper Self-Attention (Al-Rfou et al., 2018)	1.06
Transformer-XL 88M (Dai et al., 2018)	1.03
Transformer-XL 277M (Dai et al., 2018)	**0.99**
Sparse Transformer 95M (fixed)	**0.99**
ImageNet 64x64	
PixelCNN (Oord et al., 2016)	3.57
Parallel Multiscale (Reed et al., 2017)	3.7
Glow (Kingma & Dhariwal, 2018)	3.81
SPN 150M (Menick & Kalchbrenner, 2018)	3.52
Sparse Transformer 152M (strided)	**3.44**
Classical music, 5 seconds at 12 kHz	
Sparse Transformer 152M (strided)	**1.97**

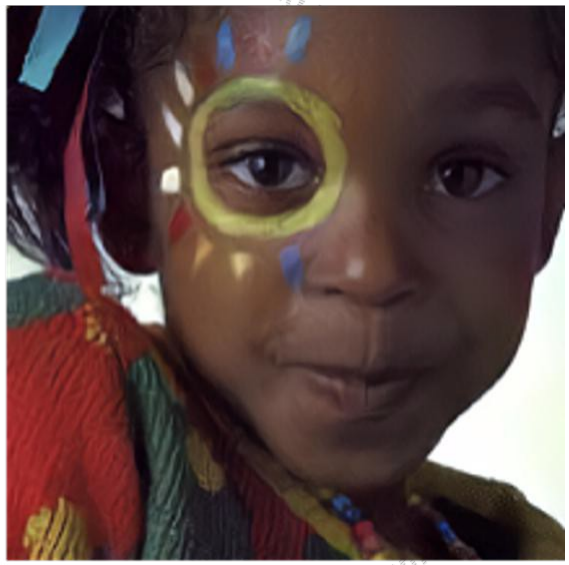
Generative models provide better bit-rates than distribution-unaware compression methods like JPEG, etc.

Rewon Child, Scott Gray, Alec Radford, Ilya Sutskever, Generating Long Sequences with Sparse Transformers,

Compression - Lossy

JPEG

JPEG2000



WaveOne

Oren Rippel, Lubomir Bourdev, Real-Time Adaptive Image Compression, ICML 2017

Downstream Task – Sentiment Dectection

This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot, Munro, and Amy are beautifully developed and their interactions are exciting, complex, and fast-paced throughout this impressive novel. And about 99.8 percent of that got lost in the film. Seriously, the screenplay AND the directing were horrendous and clearly done by people who could not fathom what was good about the novel. I can't fault the actors because frankly, they never had a chance to make this turkey live up to Crichton's original work. I know good novels, especially those with a science fiction edge, are hard to bring to the screen in a way that lives up to the original. But this may be the absolute worst disparity in quality between novel and screen adaptation ever. The book is really, really good. The movie is just dreadful.

[Radford et al., 2017]

Downstream Tasks - NLP (BERT Revolution)

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
	2	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
+	3	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5	49.1
+	4	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
	5	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
	6	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
	7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+	8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5	47.9
+	9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
+	10	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
+	11	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
	12	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
	13	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+	14	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	15	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

https://gluebenchmark.com/leaderboard

Downstream Tasks - Vision (Contrastive)

Method	Architecture	mAP
Transfer from labeled data:		
Supervised baseline	ResNet-152	74.7
Transfer from unlabeled data:		
Exemplar [17] by [13]	ResNet-101	60.9
Motion Segmentation [47] by [13]	ResNet-101	61.1
Colorization [64] by [13]	ResNet-101	65.5
Relative Position [14] by [13]	ResNet-101	66.8
Multi-task [13]	ResNet-101	70.5
Instance Discrimination [60]	ResNet-50	65.4
Deep Cluster [7]	VGG-16	65.9
Deeper Cluster [8]	VGG-16	67.8
Local Aggregation [66]	ResNet-50	69.1
Momentum Contrast [25]	ResNet-50	74.9
Faster-RCNN trained on CPC v2	ResNet-161	**76.6**

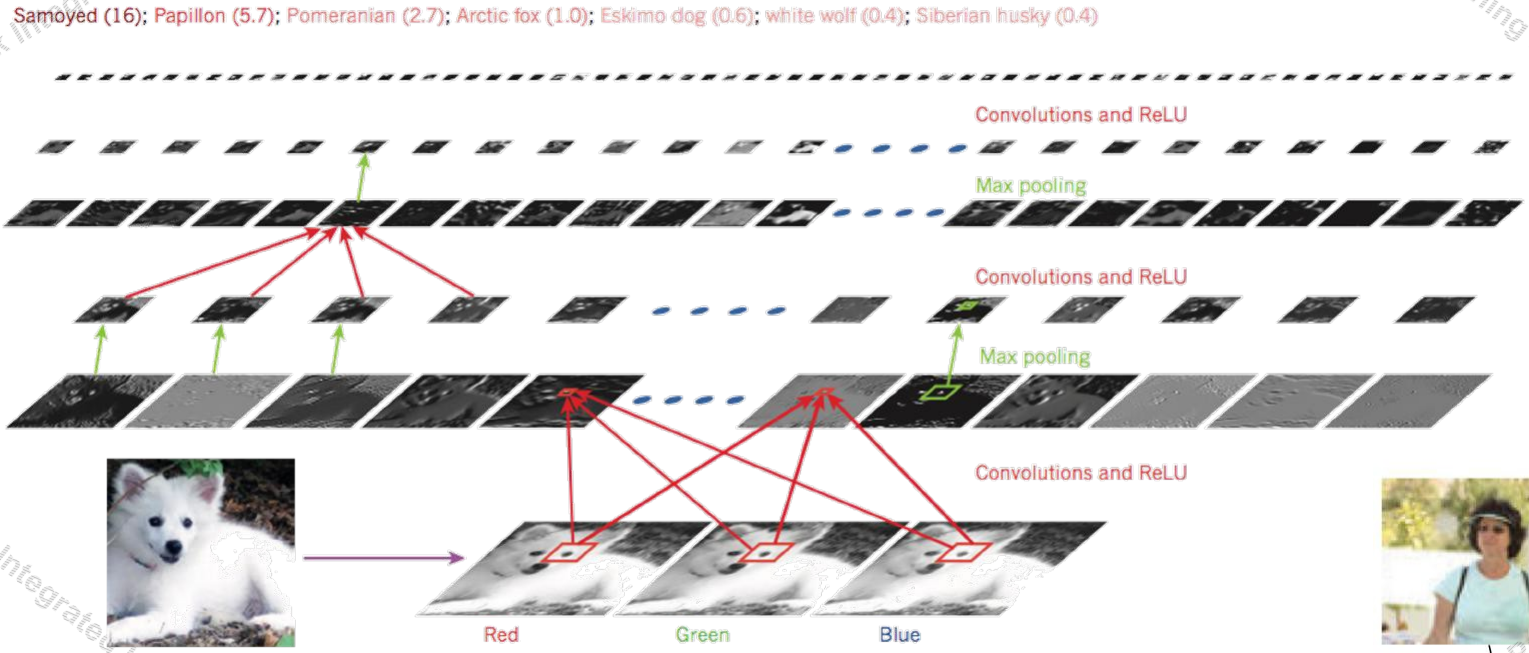
people.eecs.berkeley.edu
The Gelato Bet
Bets used to be a thing in scientific circles in days past. In oxbridge senior common rooms you can still find old betting books where bets between the dons are recorded; it makes for very amusing reading. At Berkeley, we try to uphold this tradition, except that instead of smoke-filled common rooms, we do it at the (now sadly defunct) Cafe Nefeli. The following was one such bet, made on Sept 23, 2014, hands shaken in front of three bemused witnesses (Kateria Fragkiadaki, Philipp Krähenbühl, and Georgia Gkioxari, see photo):
"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (2 scoops: one chocolate, one vanilla)."
The back story of the bet is as follows. R-CNN came out in CVPR 2014 with really impressive results on PASCAL VOC detection. I think this was a key moment when the more sceptical members within the computer vision community (such as myself) finally embraced deep learning. However, there was a complication: PASCAL VOC was said to be too small to train a ConvNet from scratch, so the network had to be pre-trained on ImageNet first, and then fine-tuned on PASCAL. This to me felt very strange: PASCAL and ImageNet were such different datasets, with completely different label sets and biases... why would training on one help the other? During that afternoon coffee at Nefeli, I suggested that maybe the network didn't actually need the ImageNet labels, just the ImageNet images to pre-train. Basically, the scientific question I wanted answered was: does one need semantic supervision to learn a good representation? Thus, the Gelato Bet was born. To entice other reserachers to get involved, I promised to share my winning gelato with any team that will help me win the bet.
Of course, I lost. Even now, five years later, we still don't have anything that beats ImageNet pre-training for PASCAL VOC (although several methods come tantalizingly close). Indeed, the whole premise that pre-training is needed for PASCAL in the first place might be erronious. On the other hand, the bet probably played a role in getting what we now call self-supervised learning started around ICCV'15. Finally, this taught me a valuable lesson: think twice before betting against your own advisor!
Alyosha Efros
Berkeley, CA
March 2019

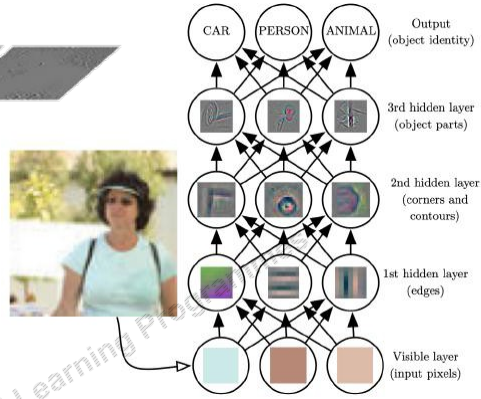
Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, Aaron van den Oord, Data-Efficient Image Recognition with Contrastive Predictive Coding, ICML 2020

Summary

- Unsupervised Learning: Rapidly advancing field thanks to compute; deep learning engineering practices; datasets; lot of people working on it.
- Not just an academic interest topic. Production level impact [example: BERT is in use for Google Search and Assistant].
- What is true now may not be true even a year from now [example: self- supervised pre-training was way worse than supervised in computer vision tasks like detection/segmentation last year. Now it is better].
- Language Modeling (GPT), Image Generation (conditional GANs), Language pre-training (BERT), vision pre-training (CPC / MoCo) starting to work really well. Good time to learn these well and make very impactful contributions.
- Autoregressive Density Modeling, Flows, VAEs, GANs, Diffusion Models.

Neural building blocks: CNNs

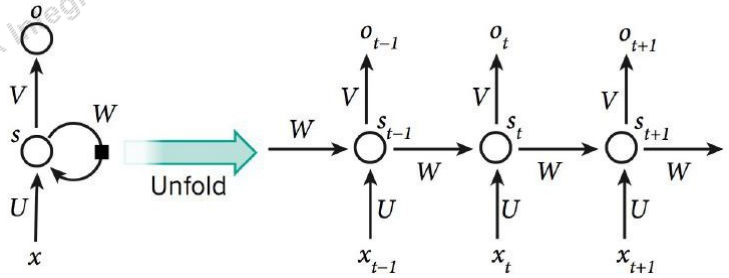




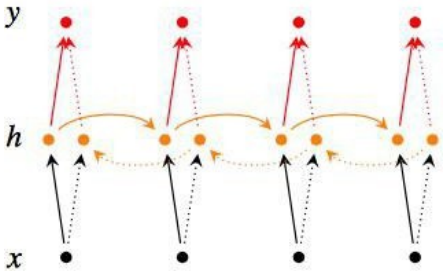
A Convolutional Neural Network (CNN)

Y. LeCun, Y. Bengio, G. Hinton, "Deep Learning", Nature, Vol. 521, 28 May 2015

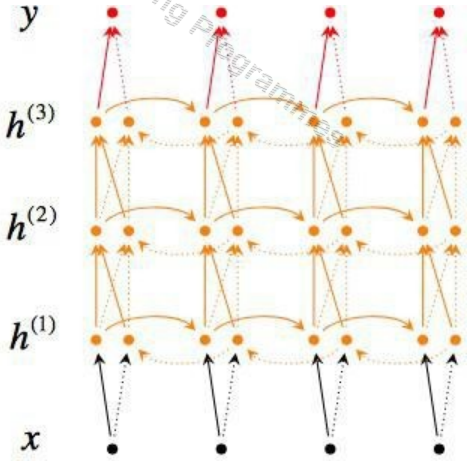
Neural building blocks: RNNs



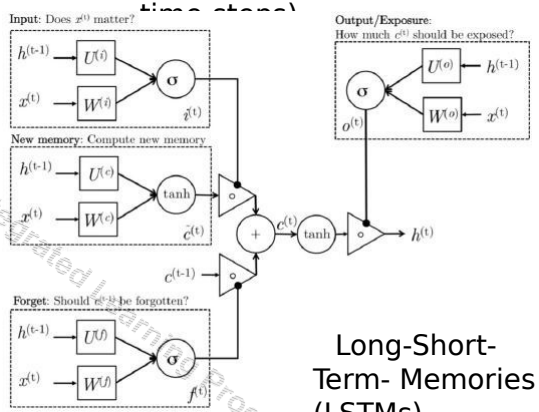
A Recurrent Neural Network (RNN) (unfolded across time steps)



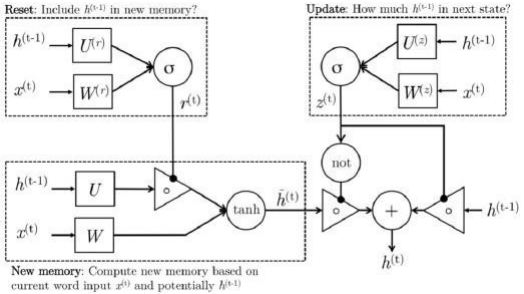
A bi-directional RNN



A deep bi-directional RNN



Long-Short-Term- Memories (LSTMs)


Reset: Include $h^{(t-1)}$ in new memory?
Update: How much $h^{(t-1)}$ in next state?
not
tanh
New memory: Compute new memory based on current word input $x^{(t)}$ and potentially $h^{(t-1)}$

Gated Recurrent Units (GRUs)

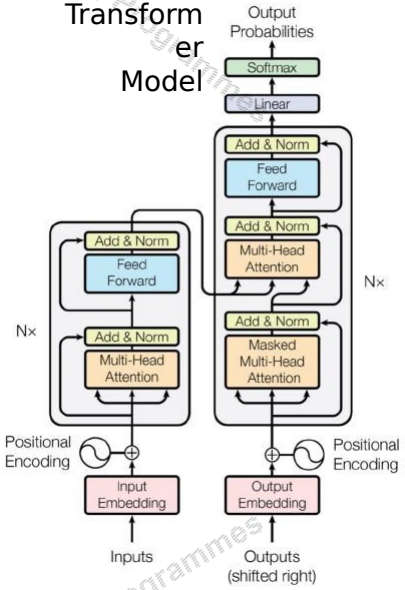
C. Manning and R Socher, Stanford CS224n Lecture 8

Y. LeCun, Y. Bengio, G. Hinton, "Deep Learning", Nature, Vol. 521, 28

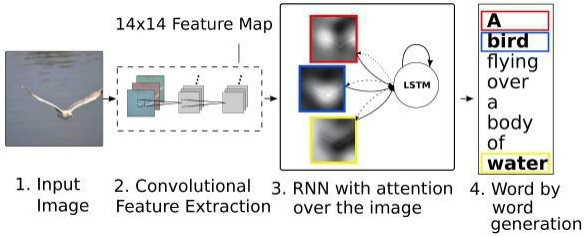
Neural building blocks: Attention mechanisms, Transformers

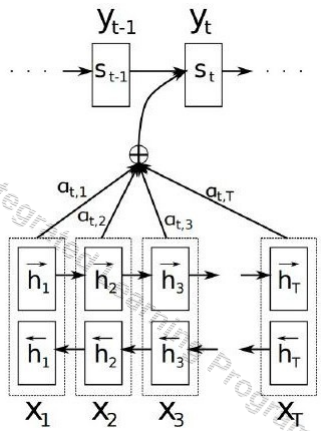
A little girl sitting on a bed with a teddy bear.
A group of people sitting on a boat in the water.
A giraffe standing in a forest with trees in the background.

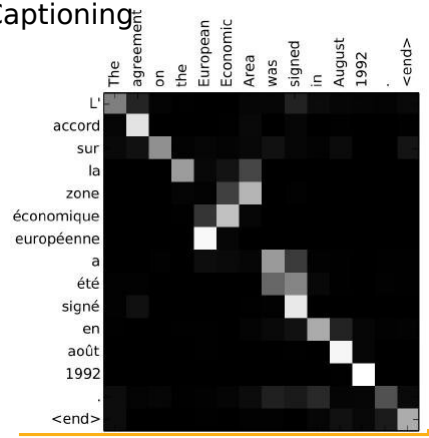
Transformer Model



Spatial Attention in Image Captioning


14x14 Feature Map
LSTM
A bird flying over a body of water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation




The agreement on the European Economic Area was signed in August 1992 . <end>
L' accord sur la zone économique européenne a été signé en août 1992 . <end>

Seq2Seq with

K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015